Escaping Datageddon

Dorothea Salo
Ryan Schryver
Graduate Support Series

Why are you here?

• You’re managing data (your own or your lab’s)
• Or you think you maybe should be
  • … maybe because your funder now says you have to?
• You’re not sure why it matters
• You’re not sure how best to do it
• You’d like to know whether you’re on the right track


Photo: Jaysin, http://www.flickr.com/photos/orijinal/3539418133/
Why manage data?

• To make your research easier!
• Because somebody else said so
  • Your lab PI
  • Your lab PI’s funder
• In case you need it later (or so as not to lose it)
• To avoid accusations of fraud or bad science
• To share it for others to use and learn from
• To get credit for producing it
• To keep from drowning in irrelevant stuff
  • ... especially at grant/project end

Research is changing...

- Research datasets were second-class citizens.
- Publications were all that mattered!
- And publishing data in print was uneconomical even when possible.
- So nobody saw anybody’s data.

- Data are now digital. The game changes!
- Data are shared more, and more openly! Open Source, Open Access, Open Data.
- There’s a lot still to be worked out about how to share, cite, credit, and license digital data.
- But data will unquestionably matter to your research careers, more than it does to your advisors’ generation.

- Learn good data habits now! You’ll need them later.

Did you know?

- Gene expression microarray data: “Publicly available data was significantly (p=0.006) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin.”
- Piwowar, Heather et al. “Sharing detailed research data is associated with increased citation rate.” PLoS One 2010. DOI: 10.1371/journal.pone.0000308
- Maybe there’s an advantage here!

Did you see?
Did you see?

News:

WisconsinView Twitter

- At 10:35AM today WisconsinView logged its 10,000th registered data user. Drop the balloons! User map: http://tinyurl.com/2chkwqv about 9 hours ago

- "Satellites See Wisconsin" is an exhibit of images on display at the airport in Madison now thru September 10 http://tinyurl.com/4mxqwvr 26 days ago

- Tonight, 9,700 people have registered for Imagery downloads at WisconsinView (since December 2005). Happy Free Data! Happy New Year! 67 days ago

- USDA 2010 NAIP county mosaics are now available for free download at WisconsinView. Sign in here: http://tinyurl.com/yaexqj6 104 days ago

...more news
Did you see?

News:

WisconsinView [Twitter]

- At 10:35AM today WisconsinView logged its 10,000th registered data user. Drop the balloons! User map: http://tinyurl.com/2chkwqv about 9 hours ago

- "Satellites See Wisconsin" is an exhibit of images on display at the airport in Madison now thru September 10 http://tinyurl.com/4mxwqvr 28 days ago

- Tonight, 9,700 people have registered for imagery downloads at WisconsinView (since December 2005). Happy Free Data! Happy New Year! 87 days ago

- USDA 2010 NAIP county mosaics are now available for free download at WisconsinView. Sign in here: http://tinyurl.com/yaexqj6 104 days ago

...more news
How to plan to keep data

Step 1: Inventory

- What data are you collecting or making?
  - Observational, experimental, simulation? Raw, derived, compiled?
  - Can it be recreated? How much would that cost?

- How much of it? How fast is it growing? Does it change?

- What file format(s)?

- What’s your infrastructure for data collection and storage like?
  - Have you ever lost data? How did that happen? Did you fix the problem?
  - How do you find it, or find what you’re looking for in it?
  - How easy is it to get new people up to speed? Or share data with others?
Step 2: Needs

- Who are the audiences for your data?
  - You (including Future You), your lab colleagues (including future ones), your PIs
  - Disciplinary colleagues, at your institution or at others
  - Colleagues in allied disciplines
  - The world!

- What are your obligations to others?
  - Funder requirements
  - Confidentiality issues
  - IP questions
  - Security

- How long do you need to keep your data?
Step 3: Process planning

• How do you and your lab get from where you are to where you need to be?
• Document, document, document all decisions and all processes!
• Secret sauce: the more you strategize up-front, the less angst and panic later.
  • “Make it up as you go along” is very bad practice!
  • But the best-laid plans go agley... so be flexible.
  • And watch your field! Best practices are still in flux.
Where to put your data

Storage, short-term

• Your own drive (PC, server, flash drive, etc.)
  • And if you lose it? Or it’s stolen? Or it breaks?
• Somebody else’s drive
  • Departmental or campus drive
  • “Cloud” drive (e.g. Dropbox)
  • Do they care as much about your data as you do? Did you read the Terms of Service? (When can they cut you off? What happens if you stop paying? Do they assert ownership or reuse rights in whatever you put there?)
• What about versioning?
• Library motto: Lots Of Copies Keeps Stuff Safe.
  • Two onsite copies, one offsite copy.
  • Keep confidentiality and security requirements in mind, of course.
Storage, long-term

- Short-term solutions are NOT GOOD ENOUGH for the long term. Sorry.
  - No, gold CD-ROMs don’t cut it either.
- If data need to persist beyond project end, you have to deal with a new kind of risk: organizational risk.
  - Servers come and go. So do labs. So do entire departments. So do you!
  - In the churn, your data may well be lost or destroyed.
  - This is especially important if you share data! Don’t let it 404!
- You need to find a trustworthy partner.
  - On campus: try the library. (Preservation is what we do!)
  - Off campus: look for a disciplinary data repository, or a journal that accepts data. (It’s a good idea to do this as part of your planning process.)
  - Let somebody else worry! You have new projects to get on with.
Things to think about

Let’s get copyright out of the way: many datasets are not copyrightable; some are.

“Facts” cannot be copyrighted in the US.

Images (including scientific images) can, as far as we know.

Beware the European database right!

You are putting yourself at great risk if you do not work out who can do what when with the data IN WRITING, FROM THE OUTSET.

All collaborators should sign off. Including (especially!) industry collaborators.

Make sure it’s okay with your funders.

If you share data, make reuse rights clear and broad.

People won’t reuse data they think they could be sued over. Would you?
File formats

• Will anybody be able to read these files at the end of your time horizon?
• Where possible, prefer file formats that are:
  • Open, standardized
  • Documented
  • In wide use
  • Easy to data-mine, transform, recast
• If you need to transform data for durability, do it now, not later.
• Fundamental question: What would someone unfamiliar with your data need in order to find, evaluate, understand, and reuse them?
• Consider the differences between someone inside your lab, someone outside your lab but in your field, and someone outside your field.
• Two parts: metadata and methods
Metadata?
What’s that?

Well,
what’s this?
Metadata?
What’s that?

Well, what’s this?

Tchaikovsky
(http://www.flickr.com/photos/alexandrovna/5508329701/)
Tania Sheka, CC-BY
Metadata is what tells you **WHAT** you’re looking at, **WHO** made it and **WHEN**, **HOW** it got there, and **WHAT** you can do with it… before you even look at the book file.
Minimal metadata to keep

The Citroen of formats?

• About the project
  • Title, people, key dates, funders and grants

• About the data
  • Title, key dates, creator(s), subjects, rights, included files, format(s), versions, checksums
  • Interpretive aids: codebooks, data dictionaries, algorithms, code

• Keep this with the data.

Methods

• Reason #1 for not reusing someone else’s data: “I don’t know enough about how it was gathered to trust it.”

• Document what you did. (A published article may or may not be enough.)

• Document any limitations of what you did.

• If you ran code on the data, document the code and keep it with the data.

• Need a codebook? Or a data dictionary?
  • If I can’t identify at sight what each bit of your dataset means, yes, you do need a codebook or data dictionary.
  • DO NOT FORGET UNITS!
Standards

• Why reinvent the wheel? If there’s a standard format for your data or how to describe it, use that!
  • Saves you time, saves the data users time.
• The tricky part is finding the right standard.
  • Standards are like toothbrushes...
  • But using standards is good hygiene!
  • Your librarian can often help you find relevant standards.
Summing up

So, these “data management plans…”

- Here’s what MIT suggests should be in them:
  - name of the person responsible for data management within your research project
  - description of data to be collected
  - how data will be documented
  - data quality issues
  - backup procedures
  - how data will be made available for public use and potential secondary uses
  - preservation plans
  - any exceptional arrangements that might be needed to protect participant confidentiality

- Feel like common sense now? Good.

Source: http://libraries.mit.edu/guides/subjects/data-management/
Help on campus

- Research Data Services: http://researchdata.wisc.edu/
  - Use the contact page!
- Your department’s liaison librarian
  - We can help you find how-tos, relevant standards, on- and off-campus archiving services, etc.
- MINDS@UW: http://minds.wisconsin.edu/
  - Data in final form that make sense as discrete files.

Photo: Jordan Pérez Nobody, http://www.flickr.com/photos/jp-/2548073841/
Thank you!

This presentation is available under a Creative Commons 3.0 Attribution license.

If you reuse it, please remember to credit the included photographs, all of which are licensed CC-BY.